Appendix 10: CAT-ASVAB Scores

Contents

Administration and Scoring of the CAT-ASVAB

During round 1 of the NLSY97, most respondents participated in the administration of the computer-adaptive form of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). The Department of Defense (DOD) used the NLSY97 sample as part of a larger effort to establish new norms for the CAT-ASVAB, a military enlistment test battery. A total of 7,127 NLSY97 respondents (or 79.3 percent of the NLSY97 sample) completed this test: 5,452, or 80.8 percent, of the cross-sectional sample and 1,675, or 74.9 percent, of the supplemental sample. In addition, young adults in two separate samples selected during the NLSY97 screening took the CAT-ASVAB during the same time period. Details about the administration of the CAT-ASVAB and the two additional samples are provided in the *NLSY97 User's Guide*.

The CAT-ASVAB consists of ten power and two speeded subtests that measure vocational aptitude in the following areas:

- Arithmetic Reasoning
- Assembling Objects
- Auto Information
- Coding Speed
- Electronics Information
- General Science
- Mathematics Knowledge
- Mechanical Comprehension
- Numerical Operations
- Paragraph Comprehension
- Shop Information

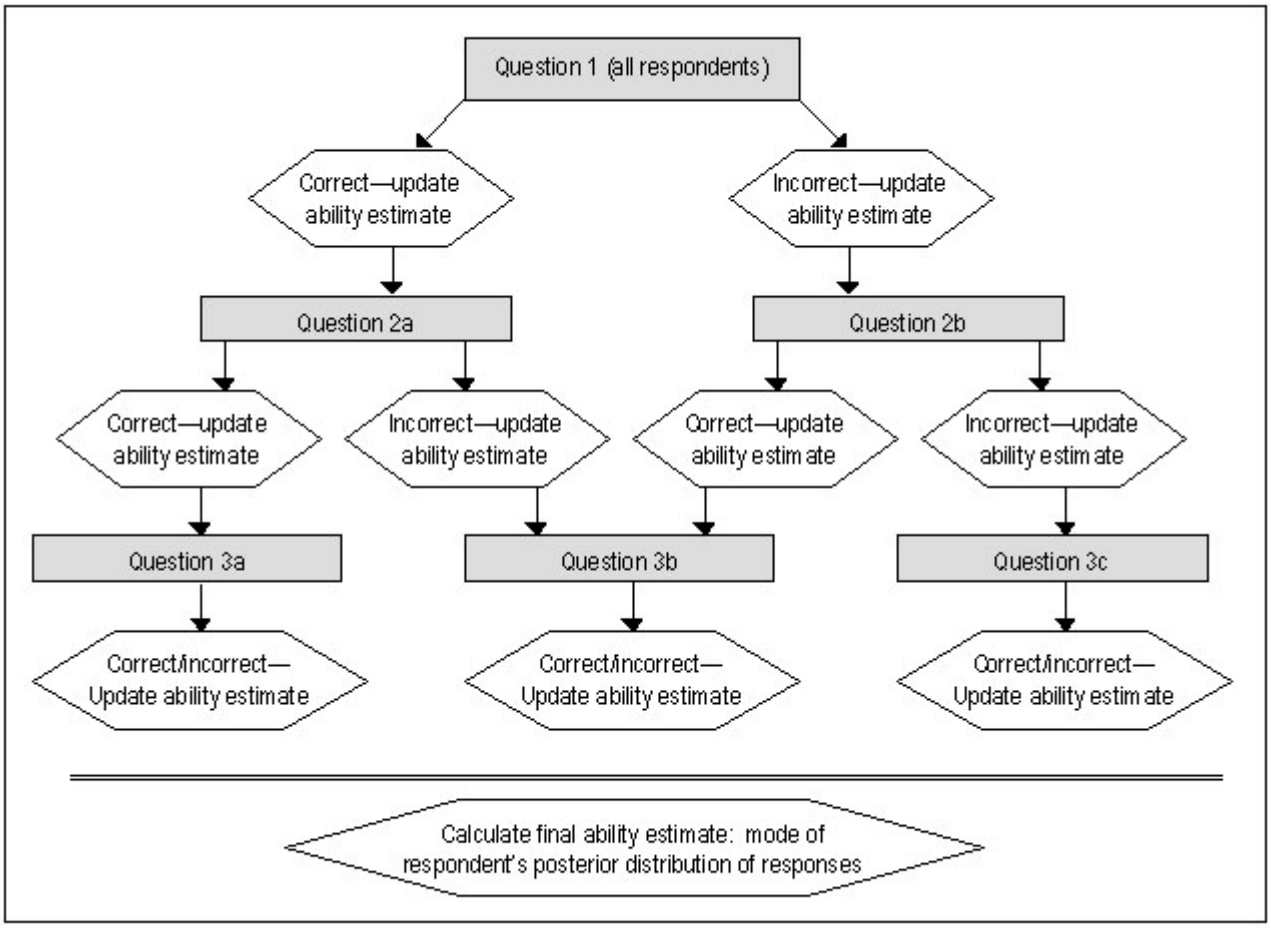
- Word Knowledge

The 10 power subtests are administered using an adaptive testing procedure that matches the difficulty level of the individual test items to the ability levels of the respondents. At the beginning of each test, the procedure selects an item of comparable difficulty for all respondents.[1] After the respondent answers the first item, the program computes a provisional estimate of ability taking into account the respondent's answer to the item. The next item is selected in accord with that provisional estimate. After the administration of each subsequent item, the provisional estimate of ability is updated using a sequential Bayesian procedure (Owen, 1969, 1975) and the updated estimate is used to match the difficulty of the next item to the ability level of the respondent. The process continues until the respondent has answered a fixed number of items (Segall et. al., 1997).[2]

Adaptive testing procedures are desirable because they require fewer items, and thus less testing time than conventional tests to obtain ability estimates of comparable precision. This is because items matched to the ability level of a respondent provide more information about the respondent's ability level.

Figure 1 is an illustration of the way an adaptive procedure might work in a simple test where each respondent answers only three questions. **Please note that this is only an example; the actual administration of the CAT-ASVAB is more complex and is not represented here.**

Figure 1. Three-Question Test Administered Using an Adaptive Testing Procedure



Because respondents do not answer the same set of items, conventional scores such as number or percent correct cannot be used to compare their performance. Instead, the program computes an item response theoretic (IRT) estimate of ability that summarizes each respondent's overall performance on the subtest. These scores are computed on a comparable scale and thus, can be compared across respondents-that is, a lower score indicates poorer performance, and a higher score indicates better performance. The final ability estimates are similar to the provisional ability estimates calculated after each test question is answered. Both calculations use IRT estimates of the item parameters obtained in a previous calibration of the CAT-ASVAB items (Segall, Moreno, and Hetter, 1997). However, the final ability estimate is not simply the last update of the provisional ability estimate. It is possible that two respondents who answered the same set of items, with the same correct or incorrect responses, would have slightly different provisional ability estimates because the items were answered in a different order. The final ability estimate takes this question order effect into account. In statistical terms, the ability scores are maximum a posterior (MAP) or Bayes modal estimates of ability. They are estimates of the mode of the posterior distribution of ability for the respondent, given the respondent's pattern of correct and incorrect responses to the subset of items completed. Final ability estimates are sometimes referred to as theta scores.

The variance of respondents' posterior distributions provides information on the precision of the ability estimates. The square root of the posterior variance, called the posterior standard deviation, is similar in meaning to the standard error of measurement and can be used to construct confidence bands on the individual estimates of ability.

The two speeded tests in the CAT-ASVAB, Coding Speed and Numerical Operations, are administered in a non-adaptive format-all respondents answer the same items in the same order. The final ability estimate for these subtests is a rate score based on the proportion of correct responses corrected for guessing, divided by the mean screen presentation time for the items (Segall et al, 1997). The mean presentation time depends on the response latencies of the respondents-that is, how long it takes them to complete the items. The rate score provides a measure of the speed and accuracy of responding.

Researchers interested in more information about test administration and scoring using Item Response Theory may wish to refer to Lord (1980).

Two additional sets of scores are scheduled for release in a future NLSY97 round. The Department of Defense will create normed scores for each section of the ASVAB. These differ from the final ability estimates in that they take into account respondent characteristics such as age and tell how a respondent performed relative to a national sample of other respondents with similar characteristics. Finally, a composite score derived from select sections of the CAT-ASVAB will provide an Armed Forces Qualifications Test score (AFQT), a general measure of trainability and a primary criterion of eligibility for Armed Forces enlistment.

User Notes: The ability estimate scores **cannot** be used in analyses in the same way as normed scores or an AFQT score. Researchers who wish to use these ability estimates should consult an expert in Item Response Theory test administration.

CAT-ASVAB Data in the NLSY97 Data Set

Included for the first time in the round 3 NLSY97 Event History data set, the ASVAB scores and related variables permit users to compare the ability levels of NLSY97 respondents in relation to one another. All ASVAB variables can be located by looking in the "Aptitude Tests" area of interest or by searching for question names that start with "ASVAB." Users should note that, although the survey year is listed in the data set as 1999 (round 3), the ASVAB was actually administered during the round 1 field period. The 1999 listing refers to the fact that the variables were added to the data set in round 3. There are five groups of variables available.

First, a variable for each subtest indicates the number of items the respondent answered in that section of the test. The number of items completed is one factor in the creation of the final ability estimates for the two speeded subtests.

The final ability estimates for each subtest comprise the second set of ASVAB variables. These estimates may have positive or negative values because the scores are on the scale of the original calibration study, which set the mean of the latent ability distribution to 0 and the standard deviation to 1 in the calibration population of respondents. Because negative codes are reserved for missing data in the NLSY97 data set, one variable cannot contain both positive and negative scores. Therefore, the final ability estimates are reported in two separate variables, one for positive scores and one for negative scores. Each respondent will have a valid value for only one of the two variables. Researchers must combine the variables to examine the scores for the full set of respondents.

User Notes: As described above, the score distribution for the final ability estimates is based on the

scale of the original calibration sample, which was set to a mean of 0 and a standard deviation of 1 in that population of respondents. The ability estimates do not have a mean of 0 in the NLSY97 sample of respondents. Overall, the ability levels of the NLSY97 respondents tended to be lower than those of the calibration sample, which was comprised of older respondents. Users will therefore notice a significantly higher number of negative scores than positive scores. However, the final ability estimates can still be used to rank the relative ability of NLSY97 respondents.

The third set of ASVAB variables, the posterior variances, provides estimates of the precision of the ability scores. These variables are available only for the 10 power subtests.

The fourth set of variables in the data set reports respondents' answers to the on-line questionnaire. These questions, mostly asked after the respondent had completed the test, collected information about the respondent's background and the testing conditions. Respondents first answered questions about their school experiences, such as the highest grade they had completed and the highest degree attained, their average grades in their last year of school, the subjects they had taken in school, and the quality of teaching in their high school science and shop classes. With respect to respondents' backgrounds, variables report respondents' ethnicities, whether English was their primary language, whether another language was spoken at home, and whether their parents worked for pay during the respondent's childhood. Finally, questions related to the ASVAB administration included whether the respondent had taken the test before, whether he or she had served in the military, the extent of the respondent's computer use prior to test administration, the main reasons the respondent took the test, and whether the respondent did as well as possible. Respondents were also asked to evaluate the comfort and noise level of the test-taking environment and to state whether the test's instructions were clear.

Finally, the data set includes item latencies (or question timings) for each of the items in the on-line questionnaire. These variables, identified by variable titles beginning with "IL," essentially measure how long the question appeared on the computer screen.

User Notes: Respondents who did not take the ASVAB are assigned a -4, valid skip, in the data. For most NLSY97 variables, a valid skip indicates that the respondent was not supposed to be asked a particular question. However, all respondents were eligible for the ASVAB administration, and a valid skip for these variables means that the respondent chose not to participate.

References

Lord, F.M. *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum, 1980.

Owen, R.J. *A Bayesian approach to tailored testing* (RB-69-92). Princeton, NJ: Educational Testing Service, 1969.

Owen, R.J. "A Bayesian sequential procedure for quantal response in the context of adaptive mental testing." *Journal of the American Statistical Association* 70 (1975): 351-56.

Segall, D.O., Moreno, K.E., Bloxom, B., and Hetter, R.D. Psychometric procedures for administering CAT-ASVAB. In W.A. Sands, B.K. Waters, and J.R. McBride (eds.), *Computerized adaptive testing:*

From inquiry to operation. Washington, DC: American Psychological Association, 1997.

Segall, D.O., Moreno, K.E., and Hetter, R.D. Item pool development and evaluation. In W.A. Sands, B.K. Waters, and J.R. McBride (eds.), *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association, 1997.

---

[1] The youngest NLSY97 respondents, those born in 1983 and 1984, were administered an "easy" form of the test in which the first question had a lower level of difficulty.

[2] In each subtest, a few respondents answered less than the fixed number of questions, presumably because they stopped taking the test or they reached the maximum time allowed for that section. These respondents still have final ability estimates, but these scores may be less precise because they are based on fewer questions.